Sources of Evidence for Neural Representation

Tyler Brooke-Wilson (tbw@mit.edu)
Dept. of Linguistics and Philosophy, MIT
32 Vassar St., Cambridge MA

Abstract:

A crucial methodological question for cognitive neuroscience is the question of what constitutes evidence of neural representation. A number of critiques over the last decade have challenged the view that correlation alone, as measured by neural decoding, is sufficient to establish representation. In response to such critiques, correlation is often augmented by a behavioral measure, showing that the decoding accuracy of a classifier and some behavioral performance measure are themselves correlated. I argue that correlation and behavioral causation together are nevertheless still insufficient for establishing representation. Inferring the existence of a neural representation on the basis of correlation and behavior alone is liable to both false positives and false negatives. Reflection on one common theory of representation (functional homomorphism theory, proposed by King and Gallistel 2010) elucidates why correlation + behavior is insufficient and suggests more direct sources of evidence. I present this theory and explain its implications for the question of empirical evidence of representation. Along the way I draw out some of the connections between the functional homomorphism theory of representation and predictive theories of perception.

Keywords: Representation, Neural Codes, Intuitive Physics, Functional Homomorphisms, Neural Decoding.

Correlation is not representation:

One of the core aims of cognitive neuroscience is to uncover the representations that allow the brain to process information about its environment. One natural way to do so is to show subjects a selection of stimuli and determine which neural responses carry information about the stimulus presented. This broad method encompasses both the localization studies of the early days of fMRI (Kanwisher, 2000) and more recent Multi-Voxel Pattern Analysis (Cox & Savoy, 2003; Haynes, 2015). While these methods have proven enlightening, they are also subject to much legitimate criticism. In particular, many have pointed out that the correlations on which these methods are based are an imperfect guide to the representations the brain actually employs (Andersen, T. Oates, 2010; Brette, n.d.; Ritchie, Kaplan, & Klein, 2019; Todd, Nystrom, & Cohen, 2013). The central worry of these critiques is that correlations are cheap — any number of features of a stimulus might give rise to a differential response in select brain areas and allow for stimulus identity to be decoded with reasonable accuracy.

Worries about spurious correlations can sometimes be addressed by careful experimental controls, but running such controls is often labor intensive, requires strong priors on the part of the experimentalist viz. what confounds are likely, and become harder and harder as the representational contents under investigation become harder and harder to pull apart. These worries become critical when the representational contents under investigation are high level, or abstract, contents (as opposed to low level visual features such as edges or colors). This is because high level contents are likely to correlate with a number of low level visual features, not least those that are used by the visual system to infer high level contents.

To take one recent example, the physical magnitude mass, if it is visually represented, is likely to correlate with a large number of low level features (including texture, volume, and shape), as well as less obvious combinations of these and other features. Factors such as the high number of low level correlates, the need for priors over which ones to check, and the difficulty of pulling apart high level contents from the low level contents that may be used to infer them, make diagnosing representations by way experimental controls fraught. Concerns such as these and others have motivated a discussion about alternative sources of evidence for neural representation.

Correlation plus behavioral causation is not representation:

The response to these worries has been to emphasize the need to establish that the correlations detected by methods such as decoding are also behaviorally relevant (Tong & Pratte, 2011). For a neural response to constitute a representation, then it must both correlate with some aspect of the environment and be causally efficacious in supplying the larger system (the brain) with that information. Defenders of this method might show, for example, that the trials on which BOLD response or decoding accuracy is greatest tend to be those on which performance is at its best (Ritchie et al., 2019).

A growing number of studies include behavioral criteria as well as decoding in diagnosing representations**.** This is a laudable step forward. Studies which show that decoding accuracy is related to behavioral performance on a trial-by-trial basis deliver much stronger evidence of representation than do studies which use correlation alone. The mantra of correlation + (behavioral) causation is however, misleading, as diagnosing representations on this basis is liable to both false positives and false negatives.

These risks come from the many causal links that mediate perception and decision. In particular, the risk of false positives will arise when subjects use information from alternative representations which correlate with the feature of interest. For example, if mass is not visually represented, subjects might still be able to perform well on a task that requires information about the masses of visual stimuli by reasoning about other features which are visually represented and which correlate with mass, for example, visual texture (say, rock-y or paper-y). On any given occasion, correlated visual features might be sufficient both for the subject to perform well on the behavioral task and for a classifier to decode object category, passing the above test for representation even when the feature of interest is not represented.

There is a significant risk of false negatives as well. Here again the risk comes from the number of causal links between perceptual representation and a resulting decision, each of which provide opportunities for information present in perception to be lost en route to decisions. A feature which actually *is* represented in vision may fail to inform decisions for any number of reasons, including the subject's choice of strategy (which information to use in their decision) or attentional biases (which information they tend to prioritize). Worse still is that some contents which are represented may by the visual system be *in principle* inaccessible to decision processes, making them permanently immune to behavioral measures. Visual representation of high level features involves a large number of computations with attendant intermediate representations. Diagnosing the algorithms vision uses in order to recover high level features may require deciding between hypotheses about alternative intermediate representations. But intermediate representations (such as pre-constancy representations of size and color, or representations of key geometrical features of faces) are unlikely to be directly available to behavior. As such, correspondence + behavioral causation methods will necessarily miss such representations.

Better methods are needed in light of the risks of both false positives and false negatives when correspondence + behavioral causation is used to diagnose perceptual representation. A clearer understanding of what representations are will help to better assess both what counts as evidence of representation and when sufficient evidence has been gathered. In the next section I discuss one such theory of representation, the functional homomorphisms theory due to (King & Gallistel, 2010). I then discuss its application to the question of standards of evidence.

Neural Representation: Functional Homomorphism Theory

A common view in the philosophy of cognitive science holds that representation is defined by a *functional homomorphism* between brain states and features of the environment. I'll first define this notion and then illustrate the concept using face perception as an instance of high level perceptual representation for which we have a good cognitive model.

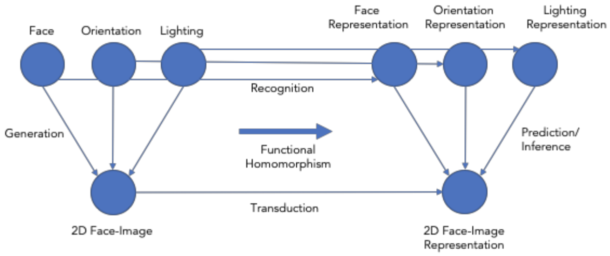
Briefly, a homomorphism is a mapping between the members of two groups, such that the structure of both groups is preserved under the mapping. That is, two structures are homomorphic if there exists some mapping f, such that $f(x) * f(y) = f(x*y)$, where * in the first case denotes an operation on members of the first group, and in the second case denotes an operation on members of the second group.

A functional homomorphism is a homomorphism between two groups such that the intra-group structure of both groups is either causal or constitutive, and the mapping between groups is causal.[1] Functional homomorphisms are interesting from the point of view of perception because they preserve probabilistic relationships among the represented features, a key idea for theories of perception that emphasize prediction or generation of the incoming stimulus (Hohwy, 2014; Yuille & Kersten, 2006).

The intuitive idea behind functional homomorphism theory is that aspects of the environment are mapped by causal processes such as transduction or pattern recognition to recurring patterns of neural activation. These patterns stand in causal relations that mirror the relations between the aspects of the environment to which they are mapped.

A recent generative model of face perception offers a quick illustration. Aspects of the environment such as the underlying geometry of a face, the lighting, and the orientation of the face collectively give rise to a 2D image of the face in question. A model by (Yildirim, 2015; Yildirim, Belledonne, Freiwald, & Tenenbaum, 2018) capitalizes on this underlying generative process and uses representations of these features (lighting, facial geometry and orientation) in order to generate a predicted image of a given face under those conditions.

On such a model, features of the environment such as lighting, facial geometry, or orientation can be inferred by way of their effects on a face image or directly observed (more realistically, inferred by another route). Similarly, a 2D projection of the face can be directly transduced or anticipated from the generating features (say in the case in which a known individual is turning towards us). This establishes a causal mapping between 2D face projection and 2D face projection image (by way of prediction or transduction) and between environmental features of facial geometry, orientation, and lighting and their representations (by way of recognition and inference). Moreover, the causal connections between facial geometry, orientation, and lighting out in the world, which generate the perceived facial projection, are mirrored by causal relations between the corresponding representations, connected by causal relations which realize processes of prediction and inference. The overall effect is the establishment of a functional homomorphism (illustrated below). On the functional homomorphism theory of representation, such relations are *definitive* of representation.



The correlation + behavioral causation methodology discussed above operates on the implicit assumption that a neural activity pattern's relationship to behavior (its behavioral efficacy) is definitive of its status as a representation. Our best theory of representation suggests otherwise. Rather, it is a putative representation's relations *to other representations* which makes it a representation with a given content. These causal inter-relations underwrite the function of representations — namely, to keep us aware of abstract or temporally or spatially unobserved parts of the world by means of inference and prediction. If this theory is right, it has broad applications to ongoing debates over the contents represented in perception.

What is the best evidence of neural representation?

Functional homomorphism theory allows us to imagine, for example, what it would be for high level physical magnitudes (such as mass), physical properties (stability and instability), and physical events (collision) to be represented in vision, as has been argued by (Battaglia, Hamrick, & Tenenbaum, 2013; Hamrick, Battaglia, Griffiths, & Tenenbaum, 2016; Scholl & Nakayama, 2002; Scholl & Tremoulet, 2000; Schwettmann, Fischer, Tenenbaum, & Kanwisher, 2018).

On the view on offer, the way to understand the neuroscientific commitments of these theories is by way of functional homomorphisms. Theories on which the visual system represents the mass of an object are theories on which (1) the brain exploits low level features such as shape to infer a representation of mass, (2) that mass representation is reliably correlated with the true mass of the object, (3) that mass representation enters into inferential relations with a number of other representations, such as volume, stability, velocity, in order to (4) causally generate low-level expectations about incoming visual stimuli (for example, the extent to which a pillow the object lands on will indent, the severity of a collision with another object, the velocity of that corresponding object post collision, etc.).

How exactly these questions should be operationalized is an interesting issue that I will not attempt to adjudicate. At first pass, however, the theory of representation on offer suggests a number of ways to more directly probe the computational relationships that are constitutive of representation. For one, if the relationships between representations are definitive of their function, then direct evidence of inferential and predictive relationships that require that content is the strongest evidence that such a content is represented.

If this is right, then decoding methods could be improved upon by aiming to provide evidence of multiple, semantically related representations. An MVPA study that shows that mass can be decoded will be more convincing if it also shows that physical magnitudes, properties, and event types to which mass is related can be decoded as well. Such studies shed light on the underlying computational structure that accounts for the representation of mass. In this way, they are less susceptible to false positives based on spurious correlations between the contents of interest and other low level contents.

Finally, to the extent that predictive theories of perception are on the right track, the signatures of representations of high level contents will be still more numerous. A visual representation of mass should be tightly related to predicted low level features, such as the degree to which a pillow is depressed by a falling object (Schwettmann et al., 2018). The ability to decode both mass and the predictions that the rational use of a mass representation would generate would lend further support to the claim that mass is represented in vision.

Conclusion:

In conclusion, correlation between a neural response and a feature of the environment, even when augmented by a behavioral measure demonstrating sensitivity to that information, is insufficient to establish neural representation. Correlation + behavioral causation regimes are susceptible to both false positives, when correlated representations are employed in subjects' decisions, and false negatives, in the case where intermediate representations are in principle unavailable to behavior.

According to our best theory of representation, the definitive features of representations are their correspondence relations *and* the mechanisms by which these relations are maintained. The mechanisms include both directly causal processes such as pattern recognition and transduction, as well as high level inferential relations. Looking for representations of high level contents holistically (i.e. in tandem with representations to which they are likely computationally related), and for the signatures of the computations they enter into, such as prediction, provides stronger evidence for representation. Methods of this kind home in on the relations that are constitutive of representation and avoid both false positives and false negatives associated with correspondence and behavioral measures alone.

References:

Andersen, T. Oates, M. (2010). A critique of multi-voxel pattern analysis. *Cognitive Science Society*.

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*(45), 18327–18332. https://doi.org/10.1073/pnas.1306572110

Brette, R. (n.d.). Is Neural Coding a Relevant Metaphor for the Brain? *Behavioral and Brain Sciences*.

Cox, D. D., & Savoy, R. L. (2003). Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. *NeuroImage*.

Haynes, J. D. (2015). A Primer on Pattern-Based Approaches to fMRI: Principles, Pitfalls, and Perspectives. *Neuron*. https://doi.org/10.1016/j.neuron.2015.05.025

Hohwy, J. (2014). *The predictive mind*. Retrieved from http://dx.doi.org/10.1093/acprof:oso/9780199682737.001.0001

Kanwisher, N. (2000). Domain specificity in face perception. *Nature Neuroscience*, *3*(8), 759–763. https://doi.org/10.1038/77664

King, A. P., & Gallistel, C. R. (2010). *Memory and the Computational Brain: Why Cogntive Science Will Transform Neuroscience*. Wiley-Blackwell.

Ritchie, J. B., Kaplan, D. M., & Klein, C. (2019). Decoding the Brain: Neural Representation and the Limits of Multivariate Pattern Analysis in Cognitive Neuroscience. *The British Journal for the Philosophy of Science*, *70*(2), 581–607. https://doi.org/10.1093/bjps/axx023

Scholl, B. J., & Tremoulet, P. D. (2000). Perceptual Causality and Animacy. *Trends in Cognitive Sciences*.

Schwettmann, S., Fischer, J., Tenenbaum, J., & Kanwisher, N. (2018). Evidence for an Intuitive Physics Engine in the Human Brain. *Cognitive Science Society*.

Todd, M. T., Nystrom, L. E., & Cohen, J. D. (2013). Confounds in multivariate pattern analysis: Theory and rule representation case study. *NeuroImage*. https://doi.org/10.1016/j.neuroimage.2013.03.039

Tong, F., & Pratte, M. S. (2011). Decoding Patterns of Human Brain Activity. *Annual Review of Psychology*, *63*(1), 483–509. https://doi.org/10.1146/annurev-psych-120710-100412

Yildirim, I., Belledonne, M., Freiwald, W., & Tenenbaum, J. (2018). *Efficient inverse graphics in biological face processing*. 1–97.

Yuille, A., & Kersten, D. (2006). Vision as Bayesian Inference : Analysis by Synthesis ? Introduction : Perception as inference. *Trends in Cognitive Sciences*, *10*(7), 301–308.